# Generative Modeling of Adversarial Lane-Change Scenario

Chuancheng Zhang<sup>1,2†</sup>, Zhenhao Wang<sup>3†</sup>, Jiangcheng Wang<sup>4</sup>, Kun Su<sup>2</sup>, Qiang Lv<sup>2</sup>, Bin Jiang<sup>1,2</sup>, Kunkun Hao<sup>4\*</sup>, Wenyu Wang<sup>2\*</sup>

Abstract—Decision-making in long-tail scenarios is crucial to autonomous driving development, with realistic and challenging simulations playing a pivotal role in testing safetycritical situations. However, the current open-source datasets do not systematically include long-tail distributed scenario data, making acquiring such scenarios a formidable task. To address this problem, a data mining framework is proposed, which performs in-depth analysis on two widely-used datasets, NGSIM and INTERACTION, to pinpoint data with hazardous behavioral traits, aiming to bridge the gap in these overlooked scenarios. The approach utilizes Generative Adversarial Imitation Learning (GAIL) based on an enhanced Proximal Policy Optimization (PPO) model, integrated with the vehicle's environmental analysis, to iteratively refine and represent the newly generated vehicle trajectory. Innovatively, the solution optimizes the generation of adversarial scenario data from the perspectives of sensitivity and reasonable adversarial. It is demonstrated through experiments that, compared to the unfiltered data and baseline models, the approach exhibits more adversarial yet natural behavior regarding collision rate, acceleration, and lane changes, thereby validating its suitability for generating scenario data and providing constructive insights for the development of future scenarios and subsequent decision training. The video demo of the evaluation process can be found at: https://www.youtube.com/watch?v=RoyfG\_B-EGw Project page: https://github.com/ChichengZZZ/ ASG

## I. INTRODUCTION

With the accumulation of autonomous driving technology, most normal driving scenarios have been widely validated and accepted as safe. However, in long-tail scenarios such as complex environments, emergencies, and extreme conditions, the lack of sufficient historical data limits the autonomous driving system's ability to respond to these situations, making it difficult to effectively predict and mitigate associated risks [1]. Lane-change, one of the most fundamental yet complex driving behaviors, is a key aspect of long-tail scenarios that challenge autonomous systems [2]. It involves dynamic, multi-vehicle interactions in two lanes. This behavior exhibits significant variability, particularly in highway and urban traffic contexts: random lane changes are more common on highways and can disrupt traffic flow, thereby reducing safety [3]; whereas forced lane changes are primarily observed in busy urban sections, potentially leading to

reduced lane capacity and generating shockwave effects [4]. In real life, lane-changing is often associated with different types of collisions, such as rear-end and side-swipe accidents. For example, in 2019, New South Wales, Australia, reported 830 lane-change collision incidents (TfNSW, 2020), and in the same year, lane-change collisions accounted for 3% of the total collision incidents in Queensland, Australia (DTMR, 2020). In the United States, side-swipe accidents constituted 13% of total collisions in 2019 (NHTSA, 2020), [5]. These statistics underscore that the risks associated with lane-changing behaviors cannot be overlooked, and that a deep understanding of lane-change decision-making and interaction processes is essential.

The growth of data volume can significantly enhance model performance; however, once the data reaches a certain volume, the growth in performance tends to plateau. Moreover, autonomous vehicle (AV) in the real world will inevitably encounter scenarios that are not present in the training data [6]. From this perspective, it becomes clear that, in the field of autonomous driving, simply expanding the data volume is not always necessary. For the vast majority of traffic scenarios, it is not essential to have an extremely large dataset to achieve coverage; instead, the focus should shift from simply expanding data to collecting targeted longtail scenario data, which is more critical to ensuring the robustness of the autonomous driving system. Thus, an idea has emerged: to refine and filter long-tail scenario data from existing open-source datasets at a refined level, and then generate these scenarios through artificial intelligence methods, thereby supplementing the data for very rare traffic scenarios. Therefore, based on the existing open-source datasets, NGSIM [7] and INTERACTION [8], we have developed a rule-based, refined approach for mining potential hazardous scenarios (see Drive Prior Module in Fig. 1). This approach employs a deep reinforcement learning (DRL) framework that simultaneously incorporates both adversarial and natural characteristics to generate highway and urban traffic flow data with candidate conditions for long-tail scenarios.

Reinforcement learning algorithms, such as Proximal Policy Optimization (PPO), have proven effective in safety-critical scenario generation [9]. To further improve the model's performance in long-tail scenarios, we innovatively enhanced the model's principles and techniques based on the PPO [10] baseline model. By introducing Leaky and Resets techniques [11], [12], we significantly increased the model's sensitivity and capacity for sustainable learning. Additionally, as shown in Fig. 1, we incorporated the Social Value Orientation (SVO) mechanism [13] to enhance the

<sup>†</sup>Both authors contributed equally to this research.

<sup>&</sup>lt;sup>1</sup>Shenzhen Research Institute of Shandong University, Shenzhen, 518057, Guangdong, China.

<sup>&</sup>lt;sup>2</sup>School of Mechanical, Electrical & Information Engineering, Shandong University, 264209 Weihai, China.

<sup>&</sup>lt;sup>3</sup>School of Mathematics and Statistics, Shandong University, 264209 Weihai, China.

<sup>&</sup>lt;sup>4</sup>Research Center of Synkrotron, Inc., Xi'an, China.

<sup>\*</sup>Corresponding author email: haokunkun@synkrotron.ai,

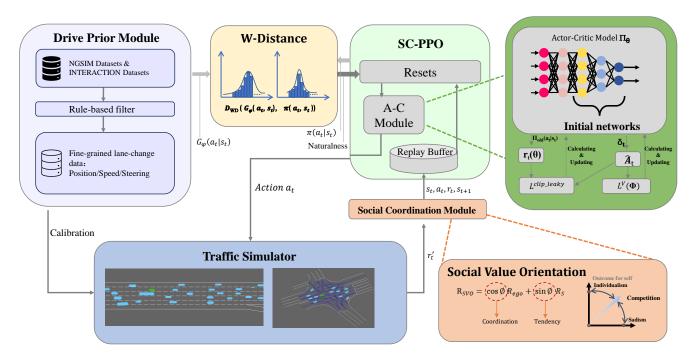


Fig. 1. The overall framework of our sensitivity and continuity scenario generation solution.

exploration capability of adversarial vehicles in scenario generation, ensuring that the generated scenario data strikes a balance between safety and naturalness [14]. Through refined data mining and model optimization for the corresponding data, our approach generates realistic long-tail scenarios more efficiently, leading to more authentic adversarial behaviors compared to baseline models. This innovative method not only provides a feasible solution for supplementing hazardous scenarios but also lays the groundwork for enhancing the safety of future autonomous driving systems. The contributions of this work have been concluded as follows.

- By selecting human driving data prone to dangerous behaviors through data mining and using reinforcement learning to generate realistic, large-scale, safety-critical scenarios, this framework serves as a foundation for testing autonomous driving systems.
- 2) To achieve a balance between naturalness and adversariality in autonomous driving scenario generation, we introduced a new reward function guided by SVO, which analyzes the influence of background vehicles on the behavior of the ego vehicle, enabling it to better understand the changes in its surroundings.
- 3) We enhanced the traditional PPO algorithm by introducing SCPPO (Sensitivity and Continuity), improving PPO's long-term learning ability and sensitivity to driving behaviors, enabling finer action exploration.

#### II. RELATED WORKS

## A. Refined Mining of Data in Existing Open Source Datasets

In recent years, with the advancement of autonomous driving research, refined data mining in open-source datasets

has become a crucial direction for enhancing model performance. Through in-depth data processing and optimization, key features in driving interaction behaviors can be more effectively captured, which in turn improves the model's ability to generalize across varied scenarios. For example, Cheng et al. [15] reduced composite errors in the nuPlan dataset using data augmentation techniques, subsequently developing a powerful baseline model. In the mining of the NGSIM dataset, Zhou et al. [16] integrated and deeply explored the NGSIM dataset using the SMARTS platform, extracting 3366 vehicle trajectories and employing PPO to train a reinforcement learning model, demonstrating superior performance in reducing hazardous events. Furthermore, Li et al. [17] first filtered the NGSIM data and then paired it with a Transformer to improve the accuracy of trajectory prediction. However, a key limitation is the simplistic nature of the data filtering process, which often overlooks the latent complexities within the data, leaving critical interactions unexplored. In contrast, Jiang et al. [18] conducted a more profound analysis of the INTERACTION dataset, extracting a dataset with high-density interaction behaviors.

## B. DRL-driven Scenario Generation

The use of DRL-based methods for generating adversarial scenarios, where adversarial agents are trained to launch attacks, has accumulated a substantial body of academic work. Sun et al. [19] employs Deep Q-Networks (DQN) to generate discrete adversarial traffic scenarios. Kuutti et al. [20] uses Advantage Actor-Critic (A2C) [21] to control a vehicle's following behavior in a surrounding vehicle scenario. Chen et al. [22] utilizes Deep Deterministic Policy Gradient (DDPG) [23] to generate adversarial strategies that

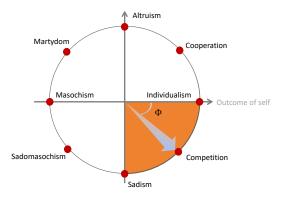


Fig. 2. The SVO ring proposed by Griesinger et al. [26]. The highlighted quadrant is used in the reward function design.

control surrounding agents to create lane-change scenarios. Wachi et al. [24] adopts Multi-Agent DDPG [25] to control two surrounding vehicles (referred to as Non-Player Characters, NPCs) to attack the ego vehicle. This approach also sets auxiliary objectives for the NPCs to avoid generating unrealistic scenarios.

### C. Social Value Orientation

Advances in behavioral and cognitive sciences have enhanced our understanding of human decision-making. Social behavior, shaped by altruism and individualism, influences drivers' actions and intentions on road. SVO reflects interpersonal traits affecting preferences in egoism, collectivism, resource allocation, and decision-making in risk-sensitive scenarios [27]. As shown in Fig. 2, Griesinger et al. [26] introduced a geometric preference model to evaluate dual choices in decomposition game experiments. In this model, the SVO is defined by the angle  $\phi$  between a straight line and the positive X-axis within the rectangular coordinate system. In the fourth quadrant, as the angle of the SVO  $\phi$  tends to  $-\pi/2$ , it reduces the favoritism towards others during driving and enhances one's competitive and antagonistic behavior.

Overall, refined mining of existing open-source datasets combined with model techniques specifically designed and tuned for them has become a prevailing trend. However, there is limited exploration regarding the refined mining of open-source datasets from multiple different scenario sources, as well as the development of DRL models tailored to integrate with these datasets based on their unique characteristics.

## III. EXPERIMENT DESIGN AND METHODOLOGY

## A. Datasets and Data Preprocessing

A refined data mining process is conducted on two widely used datasets, NGSIM and INTERACTION, to identify instances of dangerous behaviors and address the gap in long-tail distribution scenarios in existing open-source datasets.

The extraction of lane change events is based on map data and the vehicle pose sequence. For the data from NGSIM we constructed our own 5-lane highway, and for the data from INTERACTION we loaded OSM maps to obtain the network structure, then match the main vehicle's pose to the corresponding lane index in the road network. The lane

change is determined by comparing the lane indices of the current and previous frames. The conditions for a lane change are that the lane index of the current frame differs from that of the previous frame, and both indices must be within the list of nearby lanes of the main vehicle. Additionally, the corresponding frames of the lane change are recorded, and the vehicles in front and behind the main vehicle on both the original and new lanes are tracked. Algorithm 1 provides an overview of the lane change scenario extraction. Subsequently, a systematic cleaning and normalization process is applied to the extracted lane change scenario data.

# Algorithm 1 Lane Change Scenario Extraction

Input: Highway/OSM map, car trajectories

Output: Lane change scenarios

# 1) Lane change scenario pre-extraction:

Load the highway or OSM map to get the road network structure.

For each trajectory:

- 1. Get the position of the ego vehicle.
- 2. Match the vehicle position with the road network to obtain lane index.
  - 3. Record previous and current lane index.
- 4. If previous lane index differs from current lane index:
- a) Ensure both previous and current lane indices are in the nearby lane list.
  - b) Record the lane change frame.
- 5. Find the vehicles on the previous and current lanes before and after the lane change.
  - 6. Record the front and rear vehicles in both lanes.

## B. Construction of Simulation Environment

We referred to the Highway-env traffic simulation environment constructed by Hao et al. [14], and made custom adjustments to meet the specific needs of this research. The specific experimental setup and parameters are as follows:

- a) Road Structure and Environment Configuration: We built a road structure tailored based on the Highway-env platform, which includes multi-lane highways and complex intersections, as in Fig. 3.
- b) Vehicle Types and Dynamic Attributes: In the simulation environment, we introduced multiple types of vehicles, including cars, trucks, and others. Each vehicle type has different dynamic attributes, such as acceleration, maximum speed, and steering angle, to simulate the diverse driving behaviors observed in real-world traffic.

## C. Model Construction and Optimization

1) Gail-based Generation Model: Generative Adversarial Imitation Learning (GAIL) [28] is a method that combines Generative Adversarial Networks (GANs) with imitation learning. In driving behavior generation, GAIL effectively simulates the complex decision-making process of human drivers, thereby generating realistic driving behaviors.

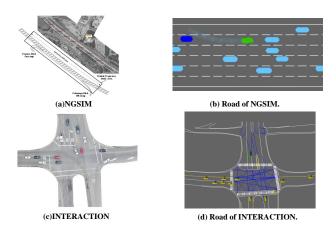


Fig. 3. Constructing road structures from the real world datasets.

Building upon the standard GAIL framework, we introduced several innovations to the PPO model to enhance its performance in complex driving scenarios:

To enhance the model's sensitivity to complex driving actions, such as lane changes on highways and intersections, the Leaky mechanism is integrated into the clipping mechanism of the PPO model, and Wasserstein Distance (W-Distance) [29], [30] is employed during training to measure naturalness of the generated behaviors. Leaky PPO allows the policy update ratio  $r(\theta)$  to maintain small gradients when exceeding a predefined threshold, preventing the vanishing gradient issue, and ensuring that the model explores the policy space more thoroughly. W-Distance, a more stable metric for measuring distribution discrepancies, effectively captures the difference between the generated policy and expert behavior, especially in high-variance or sparse reward scenarios. This enhances the model's ability to learn highrisk behaviors, such as lane changes.

Specifically, in the traditional Clipped PPO algorithm, the policy update is constrained using the ratio  $r(\theta)$  to ensure the algorithm's stability. The objective function is usually represented as follows:

$$L^{PPO}(\theta) = E_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)]$$
(1)

where  $r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$  is the ratio of the current policy to the old policy,  $\hat{A}_t$  is the advantage function, and  $\epsilon$  is the clipping threshold.

However, when the policy update ratio  $r(\theta)$  exceeds the predefined threshold  $(r(\theta) \leq 1 + \epsilon \text{ or } r(\theta) \geq 1 - \epsilon)$ , gradient information is lost, leading to an inability for the policy to further optimize. To optimize policy learning and avoid issues such as gradient vanishing, we implemented Leaky PPO, which introduces a small positive gradient when the ratio exceeds the predefined threshold. This modification preserves critical gradient information, ensuring continuous learning and enabling the model to adapt effectively to rare and challenging traffic scenarios. This modification significantly enhances the model's exploration in complex policy spaces. Additionally, Leaky PPO relaxes the ratio-based

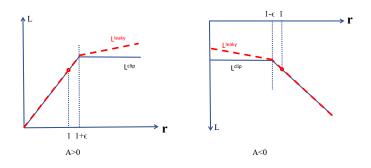


Fig. 4. This figure shows the objective function  $L_{\rm Leaky}(\theta)$  of the likelihood ratio r for the positive advantage function (left) and the negative advantage function (right). The vertical axis L represents the value of the objective function, which is used for optimizing the policy. Similarly, the red circle r=1 on each figure shows the starting point for the optimization. When the ratio r enters the saturation zone, the objective function still keeps the gradient information.

constraint, alleviating the problem of pessimistic estimation [31]. This improvement strikes a better balance between the stability of the algorithm and learning efficiency. The principle is illustrated in Fig. 4.

Specifically, the objective function of Leaky PPO consists of two parts: 1) The standard PPO loss, which is calculated using the ratio  $r(\theta)$  and the advantage function  $\hat{A}_t$ . 2) When the ratio  $r(\theta)$  exceeds the predefined threshold, a small gradient is added to prevent the vanishing gradient problem.

$$L^{LeakyPPO}(\theta) = E_t[min(r(\theta)\hat{A}_t, clip(r(\theta), l_{s,a}, u_{s,a})\hat{A}_t)]$$
(2)

Here,  $l_{s,a}$  and  $u_{s,a}$  are the new lower and upper bounds calculated based on the threshold  $\epsilon$  and the parameter  $\alpha$ , asgiven by the following formulas:

$$l_{s,a} = \alpha r(\theta) + (1 - \alpha)(1 - \epsilon) \tag{3}$$

$$u_{s,a} = \alpha r(\theta) + (1 - \alpha)(1 + \epsilon) \tag{4}$$

Where  $\alpha$  is a coefficient between 0 and 1, which controls the magnitude of the adjustment when the ratio exceeds the threshold. In the experiment, we set it to 0.01.

W-Distance, which exhibits higher robustness than traditional Kullback-Leibler (KL) divergence in handling long-tail distributions and rare events, provides a stable measurement of the difference between the generated policy and the expert behavior distribution. This effectively prevents mode collapse and enhances the naturalness of the generated data, as illustrated in Fig. 5. We use the W-Distance to calculate the naturalness reward. Specifically, for each pair of generated action  $a_1$  and expert action  $a_2$ , we compute the W-Distance through the following steps:

$$W(p_1, p_2) = \frac{1}{B} \sum_{i=1}^{B} (\| \mu_1^i - \mu_2^i \|_2^2)$$

$$+ \frac{1}{B} \sum_{i=1}^{B} \left( Tr \left( \Sigma_1^i + \Sigma_2^i - 2 \left( \sqrt{\Sigma_1^i \Sigma_2^i \Sigma_1^i} \right) \right) \right)$$
(5)

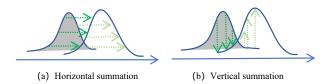


Fig. 5. (a) This section demonstrates the discrepancy between the strategies generated at the horizontal level and the distribution of expert behaviors, contributing to the prevention of mode collapse issues. (b) This section illustrates the difference between the strategies generated at the vertical level and the distribution of expert behaviors. In scenarios with sparse rewards, this approach enhances the naturalness of generated behaviors through vertical differences.

Where B is the batch size, and  $\|\mu_1 - \mu_2\|_2$  is the Euclidean distance between the mean vectors, representing the difference between the centers of the two sets of data.  $\text{Tr}(\cdot)$  denotes the trace of a matrix, which measures the difference between the covariance matrices.  $\Sigma_1$  and  $\Sigma_2$  are the covariance matrices of the two sets of data. In the above formula,  $W(p_1, p_2)$  represents the W-Distance between the generated policy and the expert behavior distribution.

The expert reward is calculated using the following formula based on the computed W-Distance loss:

$$R_{natural} = clip(\frac{\theta_w - W(p_1, p_2)}{\theta_w}, 0, 1)$$
 (6)

The range of this reward is constrained to [0,1] by  $\theta_w$ , reflecting the similarity between the generated behavior and expert behavior as quantified by the W-Distance.

To maximize the model's growth potential and prevent overfitting to early-stage data, i.e., to continuously explore new strategies during training, we introduce the environment reset mechanism (Resets). A common issue in DRL algorithms is the "prior bias" problem [32], where the agent overly adapts to early environmental interactions and neglects useful evidence from later stages, resulting in poor data quality and further hindering learning performance. To address this, a strategy is adopted of periodically reinitializing the last few layers of the neural network while retaining historical experiences in the replay buffer and updating the random seed with the current training iteration. This mechanism periodically "forgets" outdated knowledge, allowing the agent to better utilize subsequent experiences, overcome prior bias, and adapt to environmental changes. This mechanism prevents premature convergence, improving generalization, exploration in complex driving scenarios, and enhancing the model's growth potential.

With the aforementioned improvements, the GAIL-based driving behavior generation model exhibits increased sensitivity and growth potential in simulating complex behaviors, such as lane changes on highways and at intersections, significantly enhancing the realism of generated behaviors.

2) Application of Svo in The Reward Function: The incorporation of SVO into the reward function is based on the core idea of adjusting the agent's objective function so that, while optimizing its own driving behavior, it also considers

the interests of other traffic participants. This approach is particularly suitable for competitive traffic environments, where it enhances the model's adversarial capabilities while generating more rational and socially valuable driving strategies.

SVO is represented by an angle  $\phi$  that reflects the agent's behavioral preference, as depicted in Fig. 2. The closer the angle  $\phi$  is to  $-\frac{\pi}{2}$ , the more aggressive the agent becomes.

The total reward function is defined as follows:

$$L = E[R_{agent} + \lambda \cdot R_{SVO}(a_1, a_2, ..., a_n)]$$
 (7)

 $R_{\rm agent}$  represents the individual reward of the agent, measuring the impact of the agent's driving behavior on other traffic participants.  $R_{\rm SVO}$  is the constraint reward based on SVO.  $\lambda$  is a weight parameter used to balance the influence of the individual reward and the SVO reward.  $a_1, a_2, \ldots, a_n$  represent the interactions between the agent and other traffic participants. The definitions of  $R_{\rm agent}$  and  $R_{\rm SVO}$  are as follows:

$$R_{SVO} = w_1 * v_{EGO} \cdot \cos(\phi) + w_2 * \sum_{i=0}^{n} \cdot \sin(\phi)$$
 (8)

$$R_{aqent} = R_{natural} + \beta \cdot R_{adv} \tag{9}$$

where  $v_{\rm EGO}$  represents the speed of the ego vehicle.  $\phi$  is the SVO angle, indicating the agent's behavioral preference.  $w_1$  and  $w_2$  are the weight parameters for the speed term and interaction term, respectively, used to adjust the contribution of different components to the SVO reward. The term  $\sum_{i=0}^n \sin(\phi)$  represents the cumulative interaction impact of all traffic participants.  $\beta$  is the weight parameter for the adversarial reward, and  $R_{\rm adv}$  represents the adversarial reward, which aims to guide the agent to generate challenging and disruptive driving behaviors, forcing the AV to respond correctly in complex and emergency scenarios. The adversarial reward can be further divided into two parts: the distance-based reward  $r_{d,t}$  and the collision reward  $r_{c,t}$ , defined as follows:

$$R_{\text{adv}} = r_{d,t}(\mathbf{p}_{AV,t_0}, \mathbf{p}_{a,t_0}, \mathbf{p}_{AV,t}, \mathbf{p}_{a,t}) + r_{c,t}$$
(10)

$$r_{d,t}(\mathbf{p}_{AV,t_0}, \mathbf{p}_{a,t_0}, \mathbf{p}_{AV,t}, \mathbf{p}_{a,t}) = \frac{\|\mathbf{p}_{AV,t_0} - \mathbf{p}_{a,t_0}\|_2}{\|\mathbf{p}_{AV,t_0} - \mathbf{p}_{a,t_0}\|_2} - \frac{\|\mathbf{p}_{AV,t} - \mathbf{p}_{a,t}\|_2}{\|\mathbf{p}_{AV,t_0} - \mathbf{p}_{a,t_0}\|_2}$$
(11)

where  $\mathbf{p}_{\mathrm{AV},t_0}$  and  $\mathbf{p}_{\mathrm{a},t_0}$  represent the positions of the AV and the agent at the initialization time step  $t_0$ , while  $\mathbf{p}_{\mathrm{AV},t}$  and  $\mathbf{p}_{\mathrm{a},t}$  represent their positions at time step t. This metric reflects the relative proximity between the AV and the agent; the smaller the value of  $r_{d,t}$ , the closer they are, and the higher the risk of a collision. To stabilize the training process and avoid extreme rewards, a clipping function is applied, limiting the value of  $r_{d,t}$  to the range [-1,1]:

$$r_{d,t} = \text{clip}\left(\frac{\|\mathbf{p}_{AV,t_0} - \mathbf{p}_{a,t_0}\|_2 - \|\mathbf{p}_{AV,t} - \mathbf{p}_{a,t}\|_2}{\|\mathbf{p}_{AV,t_0} - \mathbf{p}_{a,t_0}\|_2}, -1, 1\right)$$
(12)

 $\begin{tabular}{ll} TABLE\ I \\ PPO\ TRAINING\ PARAMETERS \\ \end{tabular}$ 

PARAMETER	VALUE
LEARNING RATE	0.0002
BATCH SIZE	2048
NUMBER OF THREADS	2
DISCOUNT FACTOR	0.99
MAX ACTION	$[-\pi/4, 5]$
w1:w2	6:4
$\theta$ (SVO)	-45°
$\theta_w$ (W-DISTANCE)	0.9
RESET INTERVAL	1000
RESET NETWORK LAYERS	3
REPLAY BUFFER CAPACITY	100000

This clipping operation ensures that the distance metric does not become too large or too small, thereby maintaining the stability of the reward signal. In the simulation environment, collisions between the agent and the AV are crucial safety metrics. We define a collision reward  $r_{c,t}$  to reward the agent's behavior when a collision with the AV occurs. Specifically, the collision reward is calculated as follows:

$$r_{c,t} = \begin{cases} 1, & \text{if collided with the AV under test} \\ 0, & \text{if no collision} \\ -1, & \text{if collided with other vehicles} \end{cases}$$
 (13)

IV. EXPERIMENTAL PROCESS

## A. Model Training

To validate the effectiveness of the proposed driving behavior generation model, we trained it using the GAIL framework with the improved PPO algorithm. The training process involves data usage, parameter settings, and algorithm selection, as detailed below:

Different hyperparameters are set for the GAIL and PPO models during training to ensure that they could converge and perform well in complex traffic scenarios. The specific parameters for PPO are shown in Table I.

The above parameter settings ensure that the model can stably optimize during training and effectively control gradient vanishing and overfitting issues during policy updates.

The training process consists of two stages: supervised model training and natural adversarial generation model training. In the first stage, filtered samples from the original datasets are used to extract expert behavior data, crucial for training the GAIL model to simulate realistic driving behaviors. During GAIL training, the generator learns to simulate expert driving behaviors, while the discriminator enhances the generator's performance by distinguishing between generated and expert behaviors. After training, the GAIL model serves as the supervisory foundation for the PPO model, guiding the optimization process during the generation of adversarial driving behaviors.

During the PPO training phase, the model uses expert behaviors generated by GAIL as a benchmark, improving the agent's decision-making ability in complex scenarios through policy optimization. The agent interacts with the Highwayenv simulation environment, collecting samples each iteration to update the policy and value function networks. With improvements like the Leaky mechanism and W-Distance, the PPO model explores the policy space more efficiently while generating natural and diverse driving behaviors. Additionally, resetting the last three layers of the network every 1000 iterations helps avoid early overfitting and prior bias, enhancing adaptability to complex scenarios.

Through the above training process, the proposed natural adversarial generation model generates more realistic and adversarial driving scenarios, demonstrating significant performance advantages in key metrics. The above training process is simulated in the Highway-env environment and visualized through adversarial scenario examples generated by CARLA, showcasing various scenarios on highways and intersections.

## B. Experimental Metrics

1) Adversarial Reward: Adversarial reward evaluates the performance of generated driving behaviors in adversarial scenarios, serving as a key metric for optimizing individual rewards and interaction rationality in complex traffic environments. It reflects the ability to generate dangerous yet reasonable behaviors. Additionally, it combines the SVO reward with other adversarial-based reward terms. The formula is defined as follows:

$$R_{adversarial} = R_{SVO} + \beta R_{adv} \tag{14}$$

where  $R_{\rm SVO}$  represents the SVO reward, which reflects the agent's ability to balance individual rewards and adversarial behavior in complex interaction scenarios.  $R_{\rm adv}$  represents other adversarial-based rewards, and  $\beta$  is the weight parameter for the adversarial reward.

2) Dangerousness Parameter: We introduced a dangerousness parameter  $D_{\rm risk}$  to measure safety and adversarial nature of the generated model. This parameter integrates multiple key factors (Collision Rate, Distance Safety, Acceleration Stability, Trajectory Rationality) and dynamically adjusts the contribution of each metric using nonlinear functions to ensure a more accurate and reasonable evaluation. The formula for calculating  $D_{\rm risk}$  is:

$$D_r = \sum_{i=1}^4 \alpha_i \cdot f_i(R_i) + \sum_{j=1}^3 \beta_j \cdot g_j(R_j)$$

$$D_{\text{risk}} = \text{clip}\left(\frac{M - (D_r + \text{Penalty})}{M}, 0, 1\right)$$
(15)

Where  $D_{\mathrm{risk}}$  is the final dangerousness parameter, constrained within the range [0,1] by M, where values closer to 0 indicate safer behavior and values closer to 1 indicate more dangerous behavior.  $f_i(R_i)$  and  $g_j(R_j)$  are nonlinear functions designed to adjust each sub-metric's contribution. Specifically,  $f_i(R_i) = (2R_i - 1)^{\gamma_i}$  enhances the model's response to high-risk situations, and  $g_j(R_j) = \frac{1}{1+\exp(-\kappa_j R_j)}$  smooths the adjustment of metric weights, with  $\kappa_j$  controlling the sensitivity to high-risk behaviors. The penalty term

accounts for extreme driving behaviors and is given by:

Penalty = 
$$\sum_{i=1}^{n} \lambda_i \cdot (R_i - \theta_i)^{\beta_i}$$
 (16)

where  $\lambda_i$  is the penalty coefficient for each sub-metric,  $\theta_i$  is the threshold for each metric (e.g., maximum allowable acceleration, minimum safety distance), and  $\beta_i$  is the penalty strength. The dangerousness parameter provides a dynamic and comprehensive measure of the model's safety and adversarial characteristics, with values closer to 0 indicating safer behavior and those near 1 reflecting more dangerous, adversarial behavior.

## C. Experimental Results and Analysis

The experiment first evaluates the performance advantages of the proposed improved model by comparing it with several classic reinforcement learning algorithms, such as PPO, SAC, TRPO, etc. All experiments are conducted using the same training data and parameter settings to ensure the fairness of the results. The main objective of the experiment is to compare the performance of the generative model in terms of Adversarial Reward, as shown in Fig. 6.

Experimental results demonstrate that, compared to baseline models such as SAC, the SCPPO model generates more challenging driving behaviors earlier in training and maintains higher levels of adversarial reward. This suggests that the model is more effective in simulating realistic, high-risk driving scenarios, due to the high-quality data and advanced training techniques. Resets, combined with SCPPO's sensitivity and continuous learning, gives it a significant advantage over other baseline models (PPO, SAC, and TRPO) in terms of adversarial reward. This advantage is especially pronounced in generating high-risk driving behaviors, where the SCPPO model achieves higher reward levels, demonstrating a stronger ability to generate desired driving scenarios. To further validate the impact of SVO constraints on model performance, a comparative experiment is conducted to evaluate the effect of including the SVO reward. The dangerousness parameter is used to quantify

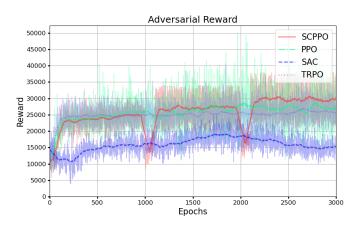


Fig. 6. Comparison with Baseline Models

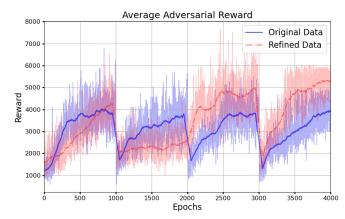


Fig. 7. Comparison of the Average Adversarial Reward Between The Refined Datasets and The Original Dataset

the effectiveness of driving behaviors generated by different models during training.

The experimental results indicate that, compared to the model without the integrated SVO reward, the SCPPO model with the integrated SVO reward exhibits a 5% increase in the dangerousness parameter when generating high-risk scenarios. The agent's behavior demonstrates more rational and adversarial characteristics. Through this comparison, we have validated the effectiveness and importance of the SVO reward in enhancing the model's ability to generate high-risk driving scenarios (such as lane changes). The model is trained using both the original datasets (NGSIM and INTERACTION) and refined datasets, with comparisons made based on metrics such as the dangerous parameters, average adversarial reward, and total adversarial reward when generating driving behaviors, as shown in Fig. 7.

The experimental results show that, as training progresses, the adversarial reward of the new dataset increases significantly more than that of the old dataset. This indicates that data selection and mining play a crucial role in enhancing the model's performance, particularly in generating adversarial driving behaviors. To verify the contribution of each improvement module to the performance of the final model, ablation experiments are designed to analyze the role of these modules and their influence on each other by removing the Leaky mechanism, Resets and the W-Distance one by one.

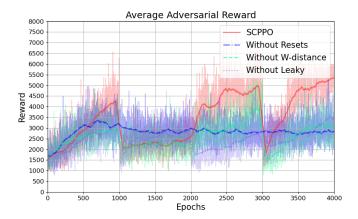


Fig. 8. Ablation Study (Comparison of Different Mechanisms: Leaky Mechanism, Resets, W-Distance)

the conclusion that the Leaky mechanism, Resets, and W-Distance are crucial for enhancing the performance of the SCPPO model. Removing any of these modules results in a decrease in adversarial reward, especially when generating high-adversarial scenarios (e.g., complex scenarios).

#### V. CONCLUSION

This study developed a rule-based refined data mining process based on existing open-source datasets. Dangerous interaction behaviors are identified in the NGSIM and IN-TERACRION datasets, and a driving behavior generation model based on the improved GAIL framework is proposed for such data, focusing on the generation of simulation data for lane-change behaviors in complex traffic scenarios. By incorporating the Leaky mechanism, W-Distance and Resets into the PPO algorithm, and integrating SVO into the reward function, the model demonstrates significant advantages in capturing and generating rare and complex driving behaviors.

Experimental results show that the proposed model outperforms the baseline models in key metrics such as dangerousness parameter, and adversarial reward, exhibiting higher sensitivity and adaptability. It is capable of generating more natural and reasonably adversarial driving behaviors based on the refined data we mined. Despite the significant results of this study, several directions remain for further exploration. For example, more refined mining of opensource datasets with additional modalities, such as Waymo and nuScenes, needs further investigation. Future work will explore incorporating image conditioning and leveraging current advances in Large Language Models (LLMs) to develop adversarial contextual reasoning generation models with small-scale production capabilities, aiming to improve the generalization of autonomous driving systems in realworld long-tail scenarios and improvesafe decision-making capabilities for handling dangerous situations.

### REFERENCES

 K. Potter ,D. Stilinski, and S. Oladimeji, "Long-Tail Learning for Rare Event Detection in Autonomous Vehicles," 2024.

- [2] W. V. Winsum, D. D. Waard, and K. A Brookhuis, "Lane change manoeuvres and safety margins," in Transportation Research Part F: Traffic Psychology and Behaviour, Vol. 2, no. 3, 1999, pp. 139-149.
- [3] A. Sasoh, and T. Ohara, "Shock wave relation containing lane change source term for two-lane traffic flow," in Journal of the Physical Society of Japan, vol. 71, no. 9, 2022, pp. 2339-2347.
- [4] B. S. Kerner, and H. Rehborn, "Experimental features and characteristics of traffic jams," in Physical review E, vol. 53, no. 2, 1996, pp. R1297.
- [5] Y. Ali, A. Sharma, and D. Chen, "Investigating autonomous vehicle discretionary lane-changing execution behaviour: Similarities, differences, and insights from Waymo dataset," in Analytic Methods in Accident Research, vol. 42, 2024, pp. 100332.
- [6] H. Li et al., "Open-sourced data ecosystem in autonomous driving: the present and future," arXiv preprint arXiv:2312.03408, 2023.
- [7] E. Leurent et al., "An environment for autonomous driving decisionmaking", 2018.
- [8] W. Zhan et al., "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," arXiv preprint arXiv:1910.03088, 2019.
- [9] K. Hao, W. Cui, L. Liu, Y. Pan, and Z. Yang, "Integrating Data-Driven and Knowledge-Driven Methodologies for Safety-Critical Scenario Generation in Autonomous Vehicle Validation," in 2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C). IEEE, 2024, pp. 970-981.
- [10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.
- [11] E. Nikishin, M. Schwarzer, P. D'Oro, P. Bacon, and A. Courville, "The primacy bias in deep reinforcement learning," in International conference on machine learning. PMLR, 2022, pp. 16828-16847.
- [12] X. Han, H. Afifi, H. Moungla, and M. Marot, "Leaky PPO: A Simple and Efficient RL Algorithm for Autonomous Vehicles," in 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024, pp. 1-7.
- [13] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. Fallah, "Social coordination and altruism in autonomous driving," in IEEE Transactions on Intelligent Transportation Systems. IEEE, vol. 23, no. 12, 2022, pp. 24791-24804.
- [14] K. Hao, W. Cui, Y. Luo, L. Xie, Y. Bai, J. Yang, S. Yan, Y. Pan, and Z. Yang, "Adversarial safety-critical scenario generation using naturalistic human driving priors," in IEEE Transactions on Intelligent Vehicles. IEEE, 2023
- [15] J. Cheng, Y. Chen, X. Mei, B. Yang, B. Li, and M. Liu, "Rethinking imitation-based planners for autonomous driving," in 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 14123-14130.
- [16] Y. Zhou, and Y. Chen, "Learning to drive in the NGSIM simulator using proximal policy optimization," in Journal of Advanced Transportation, vol. 2023, no. 1, 2023, pp. 4127486.
- [17] X. Li, J. Xia, X. Chen, Y. Tan, and J. Chen, "SIT: A spatial interaction-aware transformer-based model for freeway trajectory prediction," in ISPRS International Journal of Geo-Information, vol. 11, no. 2, 2022, pp. 79.
- [18] X. Jiang et al., "InterHub: A Naturalistic Trajectory Dataset with Dense Interaction for Autonomous Driving," arXiv preprint arXiv:2411.18302, 2024.
- [19] H. Sun, S. Feng, X. Yan, and H. Liu, "Corner case generation and analysis for safety assessment of autonomous vehicles," in Transportation research record, vol. 2675, no. 11, 2021, pp. 587-600.
- [20] S. Kuutti, S. Fallah, and R. Bowden, "Training adversarial agents to exploit weaknesses in deep control policies," in 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 108-114.
- [21] V. Mnih, "Asynchronous Methods for Deep Reinforcement Learning," arXiv preprint arXiv:1602.01783, 2016.
- [22] B. Chen, X. Chen, Q. Wu, and L. Li, "Adversarial evaluation of autonomous vehicles in lane-change scenarios," in IEEE transactions on intelligent transportation systems. IEEE, vol. 23, no. 8, 2021, pp. 10333-10342.
- [23] T. Lillicrap, "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971, 2015.
- [24] A. Wachi, "Failure-scenario maker for rule-based agent using multiagent adversarial reinforcement learning and its application to autonomous driving," arXiv preprint arXiv:1903.10654, 2019.

- [25] R. Lowe, Y. Wu, A. Tamar, J. Harb, P.r Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in Advances in neural information processing systems, vol. 30, 2017.
- [26] D. W. Griesinger, and J. W. Livingston, "Toward a model of interpersonal motivation in experimental games," in Behavioral science, vol. 18, no. 3, 1973, pp. 173-188.
- [27] P. A. Lange, and W. B. Liebrand, "Social value orientation and intelligence: A test of the goal prescribes rationality principle," in European Journal of Social Psychology, vol. 21, no. 4, 1991, pp. 273-292.
- [28] J. Ho, and S. Ermon, "Generative adversarial imitation learning," in Advances in neural information processing systems, vol. 29, 2016.
- [29] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN: Machine Learning," in Proceedings of the 34th International Conference on Machine Learning(ICML), Vol. 70, 2017, pp. 214–223.
- [30] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," in International journal of computer vision, vol. 40, 2000, pp. 99–121.
- [31] J. Markowitz, and E. W. Staley, "Clipped-Objective Policy Gradients for Pessimistic Policy Optimization," arXiv preprint arXiv:2311.05846, 2023.
- [32] M. Bauböck et al., "Modeling the orbital motion of Sgr A\*'s near-infrared flares," in Astronomy & Astrophysics, vol. 635, 2020, pp. A143.